POLSKIE BIBLIOTEKI CYFROWE 2010 • **str. 125-129**

Wykorzystanie metadanych z polskich bibliotek cyfrowych

MARCIN WERLA

Poznańskie Centrum Superkomputerowo-Sieciowe

Streszczenie

Niniejszy referat ma na celu przedstawienie możliwości wykorzystania metadanych opisujących obiekty udostępniane w polskich bibliotekach cyfrowych do tworzenia wizualnych interfejsów eksploracji zasobów tych bibliotek. Referat rozpoczyna się omówieniem istotności automatycznego wykorzystania metadanych. Następnie przedstawione są możliwości zautomatyzowanego wykorzystania istniejących metadanych temporalnych i przestrzennych, jak i propozycje działań pozwalających na poprawę stanu obecnego w tym zakresie.

Słowa kluczowe: metadane, rozproszone biblioteki cyfrowe, standaryzacja, promocja zbiorów

Wstęp

Biblioteki cyfrowe służą do publikowania on-line kolekcji obiektów cyfrowych zapisywanych w różnych formatach. Obiekty te umieszczane są w sieci razem z opisującymi je metadanymi. Obecnie, w zdecydowanej większości bibliotek cyfrowych, to właśnie metadane są podstawą do odkrywania przez użytkowników udostępnianych zasobów, zwłaszcza jeżeli zasoby te to materiały historyczne albo materiały wizualne (np. pocztówki) nie posiadające osadzonych w treści informacji tekstowych, które można by w prosty sposób przeszukiwać.

Interfejsy dostępowe bibliotek cyfrowych pozwalają na proste czy zaawansowane przeszukiwanie metadanych, ich przeglądanie oraz czasami dostarczają również bardziej atrakcyjnych narzędzi do wizualnej eksploracji, takich jak oś czasu czy prezentacja zasobów na mapie. Możliwości skonstruowania takich zaawansowanych elementów interfejsu oraz ich funkcjonalność i efektywność są ściśle powiązane z jakością metadanych dostępnych w danej bibliotece cyfrowej, w tym również ze standaryzacją ich zapisu.

Innym aspektem wykorzystania metadanych jest przekazywanie ich do zewnętrznych systemów informacyjnych. Przy takim wykorzystaniu mamy trzy możliwe scenariusze. W pierwszym obsługa biblioteki cyfrowej świadomie nawiązuje współpracę z zewnętrzną firmą czy instytucją w celu przekazywania metadanych do zewnętrznego serwisu. Przykładem mogą tu być serwisy, takie jak WorldCat, Europeana, DART-Europe czy Federacja Bibliotek Cyfrowych. Metadane są tu zazwyczaj przekazywane za pomocą wyspecjalizowanego protokołu komunikacyjnego, takiego jak chociażby OAI-PMH. Drugi scenariusz wykorzystania metadanych obejmuje sytuację, gdy zewnętrzny serwis informacyjny pozyskuje informacje o gromadzonych w bibliotece cyfrowej zasobach bez szczególnej współpracy z twórcami tej biblioteki. Tak właśnie działają wyszukiwarki internetowe, np. Google czy Yahoo, indeksując w sposób zautomatyzowany strony WWW interfejsu użytkownika biblioteki cyfrowej. W tej grupie znajdują się również wyspecjalizowane serwisy wyszukiwawcze, takie Google Scholar, które nastawione są na indeksowanie komercyjnych wydawnictw oraz instytucjonalnych repozytoriów i bibliotek cyfrowych. Serwisy takie starają się wyekstrahować z indeksowanych stron WWW metadane opisujące zawartość indeksowanych baz. Metadane te są następnie wykorzystywane np. do analizy cytowań artykułów naukowych. W trzecim scenariuszu wykorzystania metadanych w zewnętrznych systemach informacyjnych mamy do czynienia z czytelnikiem, który odwiedzając bibliotekę cyfrową znajduje interesujący obiekt i chce zachować informacje na jego temat do dalszego wykorzystania. Tutaj najpopularniejszymi narzędziami są programy, takie jak RefMan, RefWorks czy Zotero – tzw. „menadżery bibliografii".

W każdym z powyższych scenariuszy mamy problem interoperacyjności metadanych. Problem ten dotyczy, po pierwsze, uzgodnienia schematu metadanych pomiędzy biblioteką cyfrową a systemem z niej korzystającym. Nawet w przypadku serwisów działających na zasadzie Google Scholar, odpowiednie dostosowanie schematu meta-

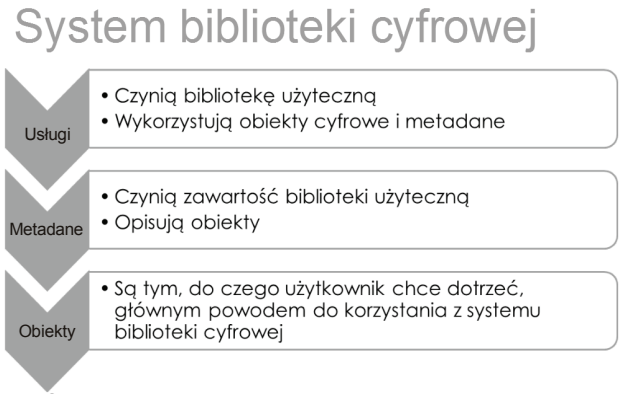
danych biblioteki cyfrowej i właściwa ekspozycja tych metadanych znacznie poprawiają stopień zaindeksowania obiektów biblioteki cyfrowej, a tym samym ich widoczność w Internecie. A to jest bardzo ważne np. dla promowania wyników prac badawczych instytucji naukowych. Kluczowy może być również format samego obiektu cyfrowego (np. w świecie publikacji naukowych dominuje format PDF).

Poza zgodnością na poziomie schematu metadanych jest jeszcze kwestia interpretacji wartości poszczególnych elementów schematu metadanych. Wykorzystanie odpowiednich standardów zapisu informacji (np. daty czy lokalizacji geograficznej) oraz użycie ogólnie uznanych klasyfikacji, kontrolowanych słowników, tezaurusów czy innych systemów typu SKOS daje szansę na osiągnięcie tzw. interoperacyjności semantycznej, czyli sytuacji, w której transmitowane informacje są czy mogą być tak samo interpretowane zarówno przez system źródłowy, jak i przez system docelowy.

Niniejszy referat ma na celu przedstawienie możliwości wykorzystania metadanych opisujących obiekty udostępniane w polskich bibliotekach cyfrowych do tworzenia wizualnych interfejsów eksploracji zasobów tych bibliotek. Przedstawione zostaną możliwości zautomatyzowanego wykorzystania istniejących metadanych temporalnych i przestrzennych, jak i propozycje działań pozwalających na poprawę stanu obecnego w tym zakresie. Ponadto przedstawione zostaną propozycje rozszerzeń schematów metadanych wykorzystywanych w polskich bibliotekach cyfrowych w celu zwiększenia możliwości wykorzystania i widoczności tych metadanych w zewnętrznych systemach, ze szczególnym uwzględnieniem współczesnych publikacji naukowych.

Rola metadanych w systemie biblioteki cyfrowej

W systemie biblioteki cyfrowej można wyróżnić trzy kluczowe i komplementarne elementy: obiekty cyfrowe, metadane oraz usługi (por. ryc. 1). Obiekty cyfrowe są tym, do czego użytkownik chce dotrzeć, są głównym powodem do korzystania z danej biblioteki cyfrowej. Metadane opisują te obiekty czyniąc w ten sposób zawartość biblioteki bardziej użyteczną. Podstawową rolą metadanych opisowych jest umożliwienie użytkownikowi odkrycia danego obiektu oraz wstępnego określenia przydatności tego obiektu, bez konieczności pobierania go. Jednak aby odkrywanie obiektów na podstawie ich metadanych było możliwe, niezbędne są specjalizowane usługi biblioteki cyfrowej. Usługi te umożliwiają eksplorację zawartości biblioteki, czyniąc ją tym samym użyteczną dla użytkownika.



Ryc. 1. Podstawowe składowe systemu biblioteki cyfrowej

Jak widać, te trzy elementy doskonale się uzupełniają i trudno jednoznacznie powiedzieć, który z tych elementów jest najważniejszy. Należy mieć jednak świadomość, że szerokie stosowanie gotowych pakietów do budowy bibliotek cyfrowych powoduje, że biblioteki te są do siebie bardzo zbliżone pod kątem oferowanych usług. To co pozostaje unikalne, to obiekty cyfrowe. To one określają charakter danej biblioteki cyfrowej i definiują podstawowe grupy użytkowników. A do zaprezentowania zawartości biblioteki cyfrowej i odnalezienia poszczególnych obiektów niezbędne są metadane.

System biblioteki cyfrowej w środowisku sieciowym

Tworząc bibliotekę cyfrową przeznaczoną do powszechnego udostępniania kolekcji za pomocą Internetu, należy mieć świadomość, że informacje publikowane w sieci są nieustannie kopiowane. Dwa podstawowe scenariusze takiego kopiowania to automatyczne kopiowanie treści do różnego rodzaju zewnętrznych serwisów oraz półautomatyczne bądź manualne kopiowanie treści przez użytkowników serwisu na własne potrzeby. Przykładami mogą być tutaj wspomniane już we wstępie roboty wyszukiwarek (wariant automatyczny) oraz użytkownicy korzystający z menadżerów bibliografii (wariant półautomatyczny bądź manualny).

Kopiowanie metadanych (zarówno automatyczne, jak i półautomatyczne) może odbywać się na dwa sposoby. Pierwszy z nich to wykorzystanie rozwiązań technicznych dedykowanych do tego typu działań. Najpopularniejszym rozwiązaniem jest tu z pewnością protokół OAI-PMH, ale warto wspomnieć również o protokole OAI-ORE, czy formacie RIS dedykowanym dla menadżerów bibliografii. Drugim sposobem, stosowanym gdy nie ma możliwości skorzystania z wygodniejszych rozwiązań, jest ekstrakcja danych z treści stron WWW. Tak najczęściej działają właśnie roboty wyszukiwarek, które wcielając się w użytkownika chodzą po stronach WWW i kopiują ich treść, a następnie dokonują jej głębokiej analizy w poszukiwaniu najistotniejszych treści.

Jak wiadomo, sieć internetowa ma charakter rozproszony. Jedną z wielu konsekwencji takiej natury sieci jest brak jednego oficjalnego punktu wejścia. Istnieją natomiast nieoficjalne, ale bardzo popularne punkty startowe, takie jak wyszukiwarka Google. Oczywiście są też wyspecjalizowane serwisy tematyczne, sprofilowane pod kątem potrzeb mniej ogólnej grupy użytkowników.

W obydwóch przypadkach serwisy będące punktami startowymi opierają swoją działalność na agregowaniu danych z innych stron (serwisów) i umożliwianiu efektywnego przeszukiwania czy przeglądania tych danych. Dlatego też kluczowe dla widoczności poszczególnych serwisów w sieci staje się pozycjonowanie, rozumiane jako zbiór działań mających na celu dobrą lokatę tych serwisów na listach wyników wyszukiwania. Jednym z kluczowych elementów pozycjonowania i zarazem widoczności zasobów biblioteki cyfrowej w sieci jest jakość i dostępność metadanych. Inne ważne elementy to dostępność obiektów cyfrowych w popularnym formacie (np. PDF) wraz z tekstową warstwą informacji oraz linkowanie do zbiorów bibliotek cyfrowych z innych serwisów.

W tym kontekście można stwierdzić, że dobre metadane to takie metadane, które są zgodne z wymaganiami twórców tych metadanych, będąc równocześnie zgodnymi z oczekiwaniami ich użytkowników, a ponadto charakteryzują się interoperacyjnością i możliwością automatycznego, dalszego wykorzystania. Wysokiej jakości metadane powinny ułatwiać użytkownikom biblioteki cyfrowej korzystanie z tej biblioteki i jej zasobów. Powinny też ułatwiać prezentację informacji o zawartości biblioteki cyfrowej w zewnętrznych serwisach, zwiększając tym samym szansę na odkrycie tej biblioteki cyfrowej przez użytkownika.

Automatyczne wykorzystanie metadanych z polskich bibliotek cyfrowych

Przy prowadzeniu i opracowywaniu nowych funkcji Federacji Bibliotek Cyfrowych kluczowe staje się pytanie, w jakim stopniu metadane dostępne w polskich bibliotekach cyfrowych nadają się do automatycznego wykorzystania do zaawansowanych sposobów eksploracji/wizualizacji zasobów tych bibliotek. Eksploracja i wizualizacja zbiorów bibliotek cyfrowych jest problematyczna nie tylko ze względu na jakość metadanych, ale również ze względu na dużą liczbę obiektów, które są udostępniane.

Korzystanie z biblioteki cyfrowej często rozpoczyna się od wyszukiwania lub przeglądania jej zbiorów w poszukiwaniu odpowiedzi na pytanie, które można w uproszczeniu zakwalifikować do jednej z czterech kategorii: „kto?", „co?", „gdzie?" oraz „kiedy?". Dwa ostatnie spośród tych pytań wiążą się oczywiście z przestrzenną i czasową interpretacją metadanych. Przykład automatycznego wykorzystania tego typu informacji do prezentacji zbiorów bibliotek cyfrowych prezentowany był w trakcie Konferencji „Polskie Biblioteki Cyfrowe" 2010. W ramach tego pokazu w czasie rzeczywistym monitorowane były publikacje wyświetlane przez użytkowników Federacji Bibliotek Cyfrowych. Metadane każdej z wyświetlonych publikacji były analizowane pod kątem określenia daty oraz miejsca wydania publikacji. Jeżeli informacje takie udawało się odnaleźć, to były one prezentowane odpowiednio na mapie oraz na osi czasu.

Na potrzeby opracowania pokazu analizie poddane zostały metadane około 320 000 obiektów cyfrowych zgromadzone w Federacji Bibliotek Cyfrowych. W celu określenia daty wydania obiektu analizowano wartości pola „Data" schematu Dublin Core, starając się doprowadzić znajdujące się tam zapisy do postaci daty zrozumiałej dla komputera. W stosunkowo krótkim czasie udało się zrobić to dla około 98% analizowanych obiektów. Głównym wyzwaniem przy opracowywaniu tego pokazu okazało się stwierdzenie, które z wartości metadanych obiektu odnoszą się do lokalizacji geograficznych. W pierwszej kolejności przeprowadzono analizę wartości wszystkich pól metadanych pod kątem występowania w nich nazw geograficznych. Jako słownik nazw wykorzystano bazę serwisu Geonames.org. Baza ta zawiera informacje na temat różnego rodzaju miejsc (nie tylko miast czy wiosek, ale również rzek, gór, itp.). Dla każdego miejsca określona jest m.in. nazwa podstawowa, szereg nazw alternatywnych, rodzaj miejsca oraz jego współrzędne geograficzne. Dla uproszczenia do eksperymentu wzięto wyłącznie tę część bazy, która dotyczy obecnego terytorium Polski. Wyniki analizy przedstawiono w tabeli 1.

Tabela 1. Wyniki analizy wartości metadanych obiektów cyfrowych w celu odkrycia powiązania z lokalizacją geograficzną (dane na podstawie próbki 319 646 obiektów)

Nazwa pola	Liczba dopasowanych publikacji	% próbki
dc:publisher	252 161	78,89
dc:subject	104 077	32,56
dc:title	10 895	3,41
dc:description	10 181	3,19
dc:creator	5 800	1,81
dc:contributor	5 357	1,68
dc:rights	3 838	1,20
dc:source	2 362	0,74
dc:date	1 026	0,32
dc:relation	886	0,28
dc:coverage	326	0,10
dc:identifier	189	0,06
dc:type	26	0,01

Jak widać zdecydowanie najwięcej nazw geograficznych występuje w polach „Wydawca" oraz „Temat". Pierwsze z tych pól zawiera bardzo często nie tylko nazwę wydawnictwa czy nazwisko wydawcy, ale również miejsce wydania. Pole „Temat" z kolei dotyczy treści publikacji i jak widać wśród znajdujących się tam słów pojawiają się również nazwy lokalizacji geograficznych. Stosunkowo dużo nazw jest też w polach „Tytuł" i „Opis". Przy analizie tytułów dość ciekawie zaznaczył się problem zmienności nazw geograficznych wraz z upływem czasu i ewolucją języka. Na skutek działania prototypowego automatycznego mechanizmu wyszukiwania powiązań publikacja starodruk zatytułowany „Podróż do Turek y Egyptu z przydanym Dziennikiem podróży do Holandyi podczas rewolucji 1787 z francuzkiego przełożona" autorstwa Jana Potockiego[1] przypisany został do miejscowości Turek w województwie wielkopolskim.

Podsumowanie

W pierwszej części artykułu pokazano, iż obecnie model biznesowy funkcjonowania bibliotek cyfrowych powinien być oparty na metadanych, które są szeroko kopiowane i wykorzystywane zarówno przez czytelników, jak i przez zewnętrzne systemy informatyczne. Metadane są obecnie narzędziem promocji biblioteki cyfrowej w sieci i powinny być jak najszerzej udostępniane z myślą o zautomatyzowanym wykorzystaniu.

[1] Publikacja dostępna on-line w CBN POLONA pod adresem http://fbc.pionier.net.pl/id/oai:www.polona.pl:27887

Podejściem docelowym, znacznie ułatwiającym takie wykorzystanie metadanych, może być stosowanie przy tworzeniu metadanych słowników czy kartotek haseł wzorcowych zawierających wykazy miejsc. W przypadku elementów, takich jak data, gdzie trudno mówić o zamkniętych słownikach, kluczowa staje się normalizacja zapisu dat – np. stosowanie notacji „rok-miesiąc-dzień". Jeżeli w jakimś konkretnym przypadku ważne jest, aby datę wyrazić w sposób słowny czy opisowy (np. „Druga Wojna Światowa"), to można podawać postać tekstową przeznaczoną dla człowieka i jako osobną wartość wprowadzać również daty w postaci przeznaczonej do interpretacji maszynowej (np. „1939/1945").

Bardzo istotne jest również odpowiednie zaprojektowanie schematu metadanych, tak aby jeżeli to możliwe, wyodrębnić określenia przestrzenne czy czasowe do dedykowanych pól schematu. Przykładem może być tutaj wyodrębnienie pola „Miejsce wydania" z pola „Wydawca" czy stosowanie pól „Zakres czasowy" i „Zakres przestrzenny", jako pól uzupełniających w stosunku do wartości pola „Temat". Wtedy znacznie łatwiejsze staje się analizowanie metadanych pod kątem wykrycia powiązań czasowych i przestrzennych.

Tego typu działania mogą wstępnie spowodować zwiększenie nakładu pracy niezbędnego na opracowanie pojedynczego rekordu metadanych, jednak ze względu na nieustannie rosnącą liczbę obiektów cyfrowych, dbałość o jakość metadanych i możliwości ich automatycznego przetwarzania wydają się nieuchronnym kierunkiem rozwoju, mającym na celu ułatwienie korzystania ze zbiorów bibliotek cyfrowych.